(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Offenlegungsschrift**

(10) **DE 196 51 788 A 1**

(51) Int. Cl.[6]: G 06 F 17/30

(21) Aktenzeichen: 196 51 788.5
(22) Anmeldetag: 12. 12. 96
(43) Offenlegungstag: 25. 6. 98

DE 196 51 788 A 1

(71) Anmelder:
Krug, Wilfried, Prof. Dr.-Ing., 01259 Dresden, DE

(74) Vertreter:
Kailuweit & Uhlemann, 01187 Dresden

(72) Erfinder:
gleich Anmelder

(56) Entgegenhaltungen:
EP 07 47 845 A1
IBM Technical Disclosure Bulletin, Vol. 38, No. 01, January 1995, S. 607/608;

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gem. § 44 PatG ist gestellt

(54) Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken

(57) Das Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken, insbesondere in regionalen und globalen Datennetzen, wie dem INTERNET, ermöglicht eine zielgenaue, zeitsparende Recherche nach beliebigen Informationen.
Durch die Verwendung intelligenter, kreativer Suchmaschinen wird die Wahrscheinlichkeit für die Ermittlung relevanter Datensätze signifikant erhöht. Die Übertragung redundanter, identischer oder ähnlicher Datensätze wird vermieden.
Durch die Lernfähigkeit der kreativen Suchmaschinen wird die Recherchengenauigkeit permanent erhöht und der Aufwand minimiert. Ebenso besteht die Möglichkeit, durch die Verwendung mehrerer, in Konkurrenz oder in Kooperation arbeitender Master-Suchmaschinen die Trefferwahrscheinlichkeit der Recherche zu erhöhen oder den zeitlichen Recherchenaufwand zu senken.

DE 196 51 788 A 1

Beschreibung

Die Erfindung betrifft ein Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken, insbesondere in regionalen und globalen Datennetzen, wie dem INTERNET.

Für die selektive Suche von Daten stehen dem Benutzer regionaler oder globaler Rechnerverbundsysteme sogenannte Suchmaschinen zur Verfügung. Dabei handelt es sich um leistungsfähige Netzwerkcomputer, auf denen Rechercheprogramme verarbeitet werden. Der Ablauf einer konventionellen Recherche geht dabei wie folgt vonstatten:

Der Benutzer gibt ein relevantes Suchwort (Deskriptor) über die Eingabeeinrichtungen (Tastatur, Maus, Spracherkennungssystem) seines Computers ein, nachdem er den Zugang zu einer beliebigen Suchmaschine innerhalb des Informationsnetzes hergestellt hat. Nachdem die Rechercheanfrage als Informationsblock an die Suchmaschine übertragen wurde, führt diese eine routinemäßige Abfrage aller mit ihr in Verbindung stehender Informationsspeicher (Server) durch. Nach Beendigung der Abfrage kann der Benutzer alle ermittelten Datensätze, in denen das relevante Suchwort (Deskriptor) gefunden wurde, in den Arbeits- oder Massespeicher seines Computers laden (kopieren).

Nachteil dieser Lösung ist, daß bei diesem Verfahren verhältnismäßig große Datenmengen ermittelt und übertragen werden, die sich bei anschließender Prüfung als nicht relevant herausstellen. Ebenso kann nicht verhindert werden, daß auf verschiedenen Servern identische Datensätze ermittelt und an den Benutzer übertragen werden. Auch hier stellt sich erst im Ergebnis einer manuellen Sichtung der ermittelten Informationen heraus, daß ein relativ hoher Anteil der ermittelten Informationen redundant erfaßt worden ist.

Neben der unerwünschten Informationsflut, die eine Auswertung brauchbarer Informationen erschwert, entstehen bei diesem Verfahren auch vermeidbare Mehrkosten durch den längeren Aufenthalt im (gebührenpflichtigen) Netz bzw. beim Zugriff auf kostenpflichtige Datenbanken.

Aufgabe der Erfindung ist es, die Nachteile des bekannten Standes der Technik zu eliminieren und ein Verfahren zu entwickeln, daß dem Informationssuchenden eine überschaubare Anzahl relevanter Datensätze selektiv bereitstellt.

Erfindungsgemäß wird die Aufgabe durch die Merkmale des kennzeichnenden Teils des Hauptanspruches gelöst. Eine alternative Lösung der Aufgabe wird im Nebenanspruch 2 vorgeschlagen. Vorzugsweise Weiterbildungen sind in den Unteransprüchen dargelegt.

Der Informationssuchende gibt – wie bisher – ein relevantes Rechercheschuwort (Deskriptor D) über die Eingabeeinrichtung seines Computers ein. Anschließend wird die Verbindung mit einer Suchmaschine hergestellt, auf der ein adaptives Rechercheprogramm abgearbeitet wird. Das Verfahren zur Informationsermittlung und -aufbereitung weist folgende Teilschritte auf:

Nachdem die Suchmaschine SM die Korrektheit und Zulässigkeit des Suchbegriffes (Deskriptor) überprüft hat (Rechtschreibprüfung, grammatikalische Prüfung, ggf. Hinweis an den Nutzer auf synonyme Bezeichnungen → Thesaurus) werden alle, mit der Suchmaschine SM in Verbindung stehenden Server $S_1 \ldots S_n$ nach diesem Deskriptor abgefragt. Von allen, auf den unterschiedlichen Servern ermittelten Datensätzen DS werden Bruchstücke (Blöcke) mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschine SM geladen. Dort wird überprüft, ob redundante Datensätze DS_{red}, gekennzeichnet durch identische Strings (Wort- und/oder Zeichenfolgen), z. B. im Titel einer wissenschaftlichen Publikation, einer Patentschrift, etc. vorhanden sind. Diese redundanten Datensätze DS_{red} werden nachfolgend gelöscht.

Parallel dazu wird die Häufung der auf den einzelnen Servern $S_1 \ldots S_n$ ermittelten Datensätze DS verglichen und die Gesamtzahl n_{ges} der ermittelten, relevanten Datensätze DS_{rel} (nach Eliminierung redundanter Datensätze) berechnet.

Ist die Anzahl der ermittelten Datensätze n_{DSrel} kleiner als ein wählbares, vom Benutzer oder vom Rechercheprogramm vorgebbares Maximum n_{DSmax} (z. B. 20 Datensätze), so wird die Recherche abgebrochen und die ermittelten Datensätze werden auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen.

Ist demgegenüber die ermittelte Anzahl relevanter Datensätze n_{DSrel} größer als dieser Grenzwert, so wird ein weiterer Recherchesuchlauf durchgeführt.

Dazu wird durch den Nutzer ein weiteres relevantes Suchwort vorgegeben. Es besteht aber auch die Möglichkeit, dem Benutzer durch das adaptive Rechercheprogramm alternative Vorschläge für weitere Deskriptoren zu unterbreiten, mit denen eine Einschränkung und Präzisierung der Recherchestrategie ermöglicht wird. So können bei der Suche nach einem technisch determinierten Schlagwort, z. B. "Kraftfahrzeug", als ergänzende Deskriptoren Suchwörter vorgegeben werden, durch die die Zweckbindung des Kraftfahrzeuges näher bestimmt wird (zum Beispiel Personenkraftwagen, Nutzkraftwagen, etc.).

Nachfolgend wird der zweite Deskriptor mit dem Deskriptor des ersten Suchlaufes additiv verbunden und der zweite Recherchendurchlauf gestartet. Dabei wird gegenüber dem ersten Recherchendurchlauf die Reihenfolge, in der die Server S_1 bis S_n abgefragt werden, nach einem Prioritätsprinzip ausgewählt. Die Prioritätsbestimmung berücksichtigt die Häufung ermittelter, relevanter Datensätze DS_{rel}, die auf den einzelnen Servern im Ergebnis des ersten Recherchendurchlaufes ermittelt wurden. Beim nachfolgenden, zweiten Recherchendurchlauf wird zunächst der Server S_i angefahren, auf dem die meisten relevanten Datensätze (bereinigt von redundanten Datensätzen) gespeichert sind. Mit absteigender Häufung (und abnehmender Trefferwahrscheinlichkeit) werden zunächst die weiteren Server abgefragt, auf denen im ersten Recherchendurchlauf relevante Datensätze ermittelt wurden.

Nachfolgend werden Bruchstücke (Blöcke) aller ermittelten, relevanten Datensätze in den Arbeitsspeicher der Suchmaschine übertragen und die Dateninhalte auf Identität (oder Ähnlichkeit) überprüft.

Redundante Datensätze werden wiederum entfernt, um die Datenmenge zu begrenzen. Neben der Identitätsprüfung kann eine Ähnlichkeitsprüfung der ermittelten Datensätze vorgenommen werden. Dabei werden Datensätze, zum Beispiel Titel von Publikationen, als ähnlich angesehen, wenn der Verfasser und das Publikationsjahr gleich sind. Ist der Autor gleich, das Erscheinungsjahr der Publikationen jedoch verschieden, so wählt das Programm den prioritätsjüngeren Datensatz aus.

Nachfolgend wird die Gesamtzahl der ermittelten, relevanten (das heißt von identischen oder ähnlichen Informationen) bereinigten Datensätze und deren Häufigkeitsverteilung auf den einzelnen angefahrenen Servern dargestellt. Ist die Gesamtzahl der ermittelten, relevanten Datensätze kleiner als ein vorgegebener Maximalwert, so wird die Recherche ab-

gebrochen und die, auf den einzelnen Servern ermittelten, relevanten Datensätze werden auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen.

Ist die Anzahl der ermittelten Datensätze nur geringfügig größer als der vorgegebene Maximalwert, so wird der Benutzer gefragt, ob er die Anzahl der Recherchenergebnisse durch einen weiteren Recherchesuchlauf mit einem gesonderten Deskriptor einschränken will oder ob er die Recherche abbrechen möchte.

Ist die Anzahl der ermittelten, relevanten Datensätze demgegenüber wesentlich größer als der gewählte Maximalwert, so wird dem Benutzer durch das adaptive Rechercheprogramm ein weiterer Deskriptor vorgeschlagen. Ebenso besteht die Möglichkeit, daß der Benutzer einen Deskriptor seiner Wahl dem nächsten Recherchesuchlauf zugrundelegt. So könnte bei der vorgehend genannten Recherche eine weitere Spezifikation des Recherchenzieles darin bestehen, daß die additiv verbundene Deskriptorenkette "Kraftfahrzeug" und "Nutzkraftfahrzeug" ergänzt wird mit dem Suchwort "Bau" oder "Straßenbau".

Die Auswahl der vom Recherchenprogramm vorgeschlagenen Deskriptoren kann sich – entsprechend dem Ziel der Recherche – an umgangs- oder fachsprachlichen Aspekten orientieren. So kann eine technisch orientierte Recherche, insbesondere eine Recherche nach relevanten Schutzrechten, unter Verwendung international üblicher Klassifikationseinteilungen (z. B. internationale Patentklassifikation IPC) vorgenommen werden.

In der vorstehend beschriebenen Weise werden iterativ bis zu n-Recherchensuchläufe durchgeführt, bis die gewünschte, maximale Anzahl relevanter Datensätze n_{DSmax} erreicht bzw. unterschritten wird. Wird im letzten Recherchendurchlauf eine Anzahl relevanter Datensätze ermittelt, die sehr klein im Vergleich zur vorgegebenen, maximal zulässigen Anzahl der Datensätze ist, so erhält der Benutzer den Hinweis, daß durch diesen letzten Recherchesuchlauf das Recherchenergebnis zu stark eingegrenzt worden ist.

Dem Benutzer wird die Möglichkeit geboten, zu den Ergebnissen des davorliegenden Recherchensuchlaufes zurückzukehren und die Recherche an dieser Stelle abzubrechen oder mit einem neuen, geänderten Deskriptor einen weiteren Recherchedurchlauf zu starten.

Eine weitere, vorteilhafte Möglichkeit zur Erhöhung der Trefferwahrscheinlichkeit einer Recherche besteht darin, daß eine Korrelation zwischen dem Datenvolumen eines Datensatzes und der Häufigkeit des Auftretens des jeweils gesuchten Deskriptors innerhalb dieses Datensatzes (Fundstelle) vorgenommen wird.

Beträgt beispielsweise das Datenvolumen eines ermittelten Datensatzes 10,0 Kilobyte (ca. 5 Seiten DIN A4) und wurde innerhalb dieses Datensatzes ein gesuchter Deskriptor (z. B. das Suchwort "Nutzkraftfahrzeug") nur einmal ermittelt, so ist die Wahrscheinlichkeit hoch, daß in dem ermittelten Datensatz lediglich peripher über "Nutzkraftfährzeuge" berichtet wird.

Die Informationsdichte I_D als Quotient aus Anzahl der ermittelten (identischen) Deskriptoren innerhalb eines Datensatzes und dem Datenvolumen (Informationsumfang, Anzahl der Seiten, etc.) dieses Datensatzes ist ein Indikator für die Wahrscheinlichkeit P_{rel}, einen relevanten Datensatz DS_{rel} zu ermitteln.

Durch das Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken wird somit eine Optimierung von Informationsrecherchen (selektive Sachrecherchen, Überblicksrecherchen, etc.) erreicht.

Eine alternative Möglichkeit zur Ermittlung einer akzeptablen Anzahl relevanter Datensätze DS_{rel} bei einer Informationsrecherche geht von der Nutzung mehrerer, vorhergehend beschriebener "kreativer" Suchmaschinen aus.

Der Informationssuchende gibt über eine Eingabeeinrichtung seines Computers ein ihn interessierendes Rechercheschlagwort (Wort, String, Zeichenkette, etc.) ein.

Der Computer stellt über die vorhandenen Kommunikationswege die Verbindung mit einer Suchmaschine SM_{Master} innerhalb des Netzwerkes her. Diese Suchmaschine stellt Ihrerseits Verbindungen mit n weiteren Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ innerhalb der Netzstruktur her. Dabei wird die Recherchenanfrage an alle angewählten Suchmaschinen weitergeleitet. Jede dieser n Suchmaschinen steht Ihrerseits mit einer Anzahl von Servern ($S_{1.1}$, $S_{1.2}$, $\ldots$ $S_{i.n-2}$, $S_{i.n-1}$ $S_{i.n}$) in Verbindung.

Im Ergebnis dieses ersten dezentralen Recherchensuchlaufes ermitteln alle n, im Slave-Modus arbeitenden, kreativen Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ eine Anzahl von Datensätzen, in denen der gewünschte Suchbegriff enthalten ist. Jede der Suchmaschinen $SM_{Slave.i}$ registriert nach Abschluß dieses ersten Recherchensuchlaufes, auf welchem der angewählten Server S_j sie welche Anzahl von Datensätzen ermittelt hat. Diese Ergebnisse werden auf einem Speicher SP der jeweiligen, im Slave-Modus arbeitenden Suchmaschine $SM_{Slave \cdot i}$ abgelegt.

In einem zweiten Verfahrensschritt erfolgt ein Vergleich der von den einzelnen Slave-Suchmaschinen $SM_{Slave.i} \ldots SM_{Slave.n}$ ermittelten Datensätze untereinander. Dabei werden wiederum redundante Datensätze ermittelt und ausgeschieden.

Nachfolgend wird die Häufigkeit der auf den einzelnen Slave-Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ ermittelten, relevanten Datensätze DS_{rel} verglichen und die Gesamtzahl $n_{ges.rel}$ der ermittelten, relevanten Datensätze DS_{rel} berechnet.

Ist die Anzahl der ermittelten Datensätze $n_{ges.rel}$ größer als ein vorgegebener oder vorgebbarer Grenzwert, so wird ein zweiter Recherchensuchlauf mit einem ergänzenden Deskriptor durchgeführt.

Die Reihenfolge, in der beim zweiten Recherchendurchlauf die Slave-Suchmaschinen $SM_{Slave.b} \ldots SM_{Slave.h}$ von der Master-Suchmaschine SM_{Master} angefahren werden, richtet sich nach der Häufung der, beim ersten Recherchensuchlauf über die einzelnen Slave-Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.i}$ ermittelten, relevanten Datensätze DS_{rel}. Wegen der höheren Trefferwahrscheinlichkeit wird so zunächst die Slave-Suchmaschine $SM_{Slave.h}$ angefahren, auf der beim ersten Recherchensuchlauf die meisten relevanten Datensätze DS_{rel} gefunden wurden. Wurde durch mehrere Slave-Suchmaschinen eine gleiche Anzahl n_{DSrel} relevanter Datensätze ermittelt, so wird als weiteres Auswahlkriterium die Informationsdichte I_D ausgewählt und danach die Reihenfolge der anzufahrenden Slave-Suchmaschinen bestimmt.

Wird bei diesem zweiten Recherchensuchlauf bereits nach der Abfrage der g-ten Slave-Suchmaschine $SM_{Slave.g}$ (wobei g<i) eine Anzahl relevanter Datensätze DS_{rel} ermittelt, die oberhalb des vorgegebenen Grenzwertes liegt, so wird die Recherche abgebrochen. Gleichzeitig wird vermerkt, welche Slave-Suchmaschinen an diesem Recherchensuchlauf nicht beteiligt waren.

Durch diese iterative Abfrage wird die Wahrscheinlichkeit, relevante Datensätze bei minimiertem Suchaufwand zu er-

mitteln, signifikant verbessert. Ein weiterer Vorteil der Einbindung einer Anzahl von n Slave-Suchmaschinen in eine Recherche besteht darin, daß die Ergebnisse einzelner Rechercheläufe temporär oder dauerhaft auf einem "Inhaltsspeicher" der beteiligten Master- oder Slave-Suchmaschine abgelegt werden können. Dieser "Inhaltsspeicher" hat vorzugsweise die Struktur einer Datenbank. Dabei wird in der Datenbank die jeweilige Recherchenanfrage (Suchwort, Deskriptor, Zeichnung, Formel, etc.) und die Anzahl der zum Rechenchenzeitpunkt temporär über diese Suchmaschine(n) in den angeschlossenen n Servern $S_1 \ldots S_n$ ermittelten, relevanten Datensätze eingetragen.

Damit besteht die Möglichkeit, bei einer späteren Recherche nach einem identischen oder begrifflich ähnlichen Suchwort (Deskriptor) qualifiziert auf die Server zuzugreifen, auf denen mit höherer Wahrscheinlichkeit relevante Datensätze abgelegt sind. Wird bei einer derartigen, zeitlich versetzen Recherche (z. B. bei einer nochmaligen Suche eines anderen Nutzers nach dem Deskriptor "Kraftfahrzeug") festgestellt, daß sich die Häufigkeitsverteilung der ermittelten, relevanten Datensätze auf den angewählten Servern geändert hat, so wird diese Drifterscheinung ebenfalls in der Datenbank "Inhaltsangabe" der jeweiligen Master- und/oder Slave-Suchmaschine gespeichert. Damit wird sichergestellt, daß bei jedem weiteren, nachfolgenden Recherchensuchlauf nach einem identischen oder inhaltsähnlichen Suchwort (Deskriptor) primär die Quellen (Server) angewählt werden, die die höchste Trefferwahrscheinlichkeit für die Ermittlung relevanter Datensätze DS_{rel} aufweisen.

Damit stellt die Datenbank "Inhaltsangabe" ein selbstlernendes System dar. So wird bei einer neuen Recherche zunächst überprüft, ob das vorgegebene Suchwort bereits identisch in der Datenbank "Inhaltsangabe" enthalten ist. Ist das nicht der Fall, so wird überprüft, ob bereits Recherchen nach ähnlichen, inhaltsgleichen Begriffen durchgeführt worden sind. Ist das der Fall, d. h. wurde beispielsweise über diese Master- oder Slave-Suchmaschine bereits eine Recherche nach dem Begriff "Kraftfahrzeug" (anstelle des ursprünglich gewählten Suchbegriffes "Nutzkraftfahrzeug") vorgenommen, so wird die Suche nach dem neuen Deskriptor "Nutzkraftfahrzeug" wegen der höheren Trefferwahrscheinlichkeit zunächst über die Slave-Suchmaschinen in den Servern durchgeführt, in denen beim letzten Recherchensuchlauf die größte Anzahl relevanter Datensätze (gegebenenfalls unter Berücksichtigung der Informationsdichte dieser Datensätze) ermittelt wurde.

Die Bewertung der Trefferwahrscheinlichkeit der in Konkurrenz arbeitenden Slave-Suchmaschinen wird dabei durch die Master-Suchmaschine vorgenommen.

Um den Zeitaufwand einer Informationsrecherche weiter zu minimieren, besteht die vorteilhafte Möglichkeit, die Rechercherfrage parallel an mehrere Master-Suchmaschinen zu leiten, die im Netzwerk autonom arbeiten oder miteinander verbunden sind. Da die Master-Suchmaschinen ihrerseits jeweils mit einer Anzahl i . . . k unterschiedlicher Server kommunizieren, werden Rechercheergebnisse ermittelt, die mit hoher Wahrscheinlichkeit repräsentativ für die untersuchte Grundgesamtheit von Informationsquellen (Servern) sind.

In einer vorteilhaften Ausgestaltung des Verfahrens zur Datenermittlung und -aufbereitung in Informationsnetzwerken werden die intelligenten, kreativen Suchmaschinen SM_{Master} über Kommunikationswege untereinander als neuronales Netz verbunden. Bei jeder Recherche werden dabei die gewonnenen Ergebnisse über die Häufigkeit ermittelter relevanter Datensätze, deren Informationsgehalt (Informationsdichte I_D) und somit die Trefferwahrscheinlichkeit auf den angefahrenen Servern protokolliert und auf ausgewählten oder auf allen, im Netz befindlichen Suchmaschinen SM_{Master} abgelegt. Durch diesen informationellen Selbstlernprozeß der Suchmaschinen SM_{Master} wird die Qualität und Ausbeute der Recherchen systematisch verbessert und der zeitliche und finanzielle Recherchenaufwand signifikant gesenkt.

Die Erfindung wird nachfolgend an einem Ausführungsbeispiel näher beschrieben.

Ein Nutzer des INTERNET in Belgien möchte sich eine Übersicht über alle Restaurants der Hansestadt Hamburg verschaffen.

Der Informationssuchende gibt über die Tastatur seines Computers als Suchstring "Restaurant Hamburg", ein. Die Obergrenze der maximal zu ermittelnden relevanten Datensätze $DS_{rel.max}$ wurde vom Informationssuchenden aus Kostengründen auf 1.000 begrenzt.

Nachfolgend wird die Verbindung des Computers mit einer Suchmaschine SM_{Master} innerhalb des Netzwerkes hergestellt. Diese Suchmaschine SM_{Master} korrespondiert ständig mit 10 Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.10}$. Die Suchmaschine SM_{Master} wählt zufällig eine Slave-Suchmaschine aus. Die im vorliegenden Fall ausgewählte Suchmaschine $SM_{SLave.6}$ korrespondiert mit 26.414 Servern weltweit.

Die Suchmaschine $SM_{SLave.6}$ sucht nunmehr in allen, mit ihr verbunden Servern nach den kummulativ auftretenden Informationen "Restaurant" und "Hamburg". Als Ergebnis wird auf dem Display des Informationssuchenden die Gesamtzahl der ermittelten Datensätze n_{ges} aufgezeigt. Die ermittelte Anzahl von 10.012 Datensätzen umfaßt alle nachgewiesenen Gaststätten in Belgien, den Niederlanden und Luxemburg mit der besonderen Etablissementbezeichnung "Hamburg".

Da dieses Recherchenergebnis nicht den Vorstellungen des Informationssuchenden entspricht, wird die Recherche an dieser Stelle nicht abgebrochen, sondern die Suchmaschine SM_{Master} wählt aus den 10, mit ihr verbundenen Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.10}$ nach dem Zufallsprinzip weitere Suchmaschinen aus, an die die Recherchenfrage weitergeleitet wird. Anschließend werden die Recherchenergebnisse übermittelt. So wurden unter Inanspruchnahme der Suchmaschine $SM_{Slave.2}$ insgesamt 2.444 Datensätze gefunden, in denen die Begriffe "Restaurant" und "Hamburg" in den USA, Kanada und Deutschland gefunden wurden.

Die Suchmaschine $SM_{Slave.1}$ ermittelte 1.436 Datensätze, in denen sich ein Hinweis auf die Suchbegriffe "Restaurant" und "Hamburg" innerhalb der Europäischen Union und in Japan findet.

Die Suchmaschine $SM_{Slave.4}$ liefert als Ergebnis 795 Datensätze, in denen Restaurants mit der Geschäftsbezeichnung "Hamburg" innerhalb der Bundesrepublik Deutschland ermittelt wurden.

Die Suchmaschine $SM_{Slave.9}$ findet bei der analogen Recherche in den, mit ihr verbundenen Servern insgesamt 1.214 Datensätze von Restaurants in Deutschland und Dänemark.

Die Suchmaschine $SM_{Slave.10}$ ermittelt insgesamt 7.117 Restaurants in Frankreich, Deutschland und den Niederlanden mit der besonderen Geschäftsbezeichnung "Hamburg".

Die Suchmaschine $SM_{Slave.5}$ findet 402 Datensätze von gleichnamigen Restaurants in der Bundesrepublik.

Die Suchmaschine $SM_{Slave.3}$ ermittelt 7.212 Gaststätten in den USA, Kanada, Mexiko, Großbritannien, Frankreich, Italien und Deutschland.

Die Suchmaschine $SM_{Slave.7}$ ermittelt 222 Datensätze, die auf Restaurants mit der besonderen Geschäftsbezeichnung "Hamburg" in den Vereinigten Staaten hinweisen.

Die Suchmaschine $SM_{Slave.6}$ findet 9.781 Datensätze mit Hinweisen auf gleichnamige Restaurants in den USA, Australien und Neuseeland.

Die Suchmaschine $SM_{Slave.8}$ ermittelt 2.006 Datensätze mit den Deskriptoren "Restaurant" und "Hamburg" mit Sitz in Großbritannien, den USA, Japan und Südkorea.

Anschließend läuft eine automatische Kreuz- und Autokorrelationsanalyse der sich im Arbeitsspeicher AS der Suchmaschine SM_{Master} befindlichen Daten ab. Dabei werden die Datensätze ermittelt, die redundant von zwei oder mehreren Suchmaschinen ermittelt wurden.

Nach Eliminierung der redundanten Datensätze werden die ermittelten, relevanten Datensätze aufgezeigt. Im vorliegenden Fall werden nur die Datensätze angesprochen, in denen die Deskriptorenkette "Restaurant, Hamburg" in Verbindung mit "Bundesrepublik Deutschland" als Resourcenquelle aufgefunden wurden. Relevante Datensätze wurden somit nur über die Suchmaschinen $SM_{Slave.1}$, $SM_{Slave.2}$, $SM_{Slave.4}$, $SM_{Slave.5}$, $SM_{Slave.9}$ und $SM_{Slave.10}$ ermittelt.

Unter Berücksichtigung der Häufung $H(DS_{rel})$ relevanter Datensätze DS_{rel} ergibt sich die Rangfolge aus dem Gesamtspektrum der Datensätze DS wie folgt:

$n_{5.rel} = 402$
$n_{4.rel} = 795$
$n_{9.rel} = 1.214$
$n_{1.rel} = 1.436$
$n_{2.rel} = 2.444$
$n_{10.rel} = 7.117$
$n_{3.rel} = 7.212$

Aufgrund der vorliegenden Obergrenze der maximal zu ermittelnden, relevanten Datensätze von $n_{relmax} = 1.000$ werden nur die Ergebnisse der Suchmaschinen $SM_{Slave.5}$ und $SM_{Slave.4}$ weiterverarbeitet.

Daneben erfolgt eine interne Bewertung aller Suchmaschinen, deren Bewertungsergebnisse in der Lernmatrix (Datenbank) der Suchmaschine SM_{Master} abgespeichert wird. In der Reihenfolge von "beste" bis "schlechteste" Suchmaschine ergibt sich folgende Reihenfolge:

$$\longrightarrow$$

$SM_{Slave.5}$, $SM_{Slave.4}$, $SM_{Slave.9}$, $SM_{Slave.1}$, $SM_{Slave.2}$, $SM_{Slave.10}$, $SM_{Slave.3}$, $SM_{Slave.6}$, $SM_{Slave.7}$, $SM_{Slave.8}$.

Die Reihenfolge wird als Wertigkeit durch eine Punktbewertung berücksichtigt. Obwohl die Suchmaschinen $SM_{Slave.6}$ ... $SM_{Slave.8}$ keine relevanten Datensätze ermittelt haben, werden diese Suchmaschinen nicht mit der Bewertungskennziffer "0" bewertet, da von diesen Suchmaschinen Datenbestände erfaßt wurden, die für eine ergänzende Recherche noch relevant sein könnten. So ist beispielsweise in dem Suchfundus "Europäische Union" Deutschland mittelbar enthalten.

Nachfolgend werden die, von den Suchmaschinen $SM_{Slave.4}$ und $SM_{Slave.5}$ ermittelten relevanten Datensätze angezeigt.

Ist der Informationssuchende mit den Ergebnissen der Recherche zufrieden, kann ein Abbruch der Recherche erfolgen.

Soll die Recherche weiter spezifiziert werden, erfolgt eine weiterer Recherchesuchlauf. Dabei wird mittels der angeschlossenen Master-Suchmaschine SM_{Master} und der mit ihr verbundenen Slave-Suchmaschinen nach der Deskriptorenkette ["Stadt" und "Hamburg" und ("Restaurant" oder "Gaststätte")] gesucht.

Prinzipiell könnte die Recherche auf die Suchmaschinen $DS_{Slave.4}$ und $DS_{Slave.5}$ beschränkt werden, die im vorigen Recherchelauf die besten Ergebnisse erbracht haben.

Ist der Informationssuchende jedoch an einer hohen Repräsentanz der ermittelten relevanten Datensätze interessiert, so werden von der Suchmaschine SM_{Master} wiederum alle 10 mit ihr korrespondierenden Suchmaschinen $SM_{Slave.1}$ bis $SM_{Slave.10}$ abgefragt. Dabei werden wegen der hohen, zu erwartenden Trefferwahrscheinlichkeit zunächst die Suchmaschine $SM_{Slave.5}$, dann die Suchmaschine $SM_{Slave.4}$, usw. angefahren.

Nach diesem zweiten Recherchenlauf werden alle ermittelten Ergebnisse aufgezeigt:
Durch die Suchmaschine $SM_{Slave.5}$ wurden 120 Restaurants in der Stadt Hamburg ermittelt. Die Datensuche über die Suchmaschine $SM_{Slave.4}$ ergab 140 Nachweise von Restaurants in der Stadt Hamburg. Die Recherche über die Suchmaschine $SM_{Slave.2}$ erbrachte 400 Datensätze von Restaurants in den USA und Deutschland mit dem Namen "Stadt Hamburg".

Im Ergebnis einer erneuten Auto- und Kreuzkorrelationsanalyse wurde festgestellt, daß alle Datensätze, die über die Suchmaschine $SM_{Slave.5}$ ermittelt wurden, sich in identischer Form im Bestand der Datensätze der Suchmaschine $SM_{Slave.4}$ befinden. Alle übrigen Datensätze sind nicht redundant. Somit bleiben 140 relevante Datensätze übrig, die dem Informationssuchenden angezeigt und in den Arbeits- bzw. Hauptspeicher seines Rechners kopiert werden, da das Abbruchkriterium $n_{rel} < n_{relmax}$ erfüllt ist. Die Trefferwahrscheinlichkeit, die von den einzelnen Suchmaschinen $SM_{Slave.1}$ bis $SM_{Slave.10}$ bei dieser speziellen Recherche erzielt wurde, wird in der Lernmatrix (Datenbank) der Suchmaschine SM_{Master} abgelegt.

Damit besteht die Möglichkeit, daß bei einer identischen oder ähnlichen Recherche eines anderen Informationssuchenden die Master-Suchmaschine SM_{Master} zunächst die Suchmaschinen $SM_{Slave.i}$ bis $SM_{Slave.m}$ auswählt, die aufgrund der bisherigen Rechercheerfahrungen die höchste Trefferquote relevanter Datensätze erwarten lassen. Kommt es

dabei aufgrund der sich ständig ändernden Datenmengen und -inhalte zu einer Verschiebung der Prioritätsliste (Rangfolge der zu erwartenden Trefferwahrscheinlichkeit), so werden auch diese Änderungen in der Lernmatrix der kreativen Suchmaschine SM_{Master} registriert, so daß eine ständige Aktualisierung erfolgt.

Ebenso werden in der Lernmatrix synonyme Deskriptoren gespeichert, die im Falle einer ergebnislosen Recherche eine Suche nach inhaltsgleichen Deskriptoren ermöglicht. Nach der "Anlernphase", in der der Aufbau der Lernmatrizen auf den Mastersuchmaschinen SM_{Master} erfolgt, erhält der Informationssuchende bei der Kontaktaufnahme mit der intelligenten Suchmaschine SM_{Master} zu Beginn seiner Sitzung eine Übersicht der recherchierbaren Fachgebiete, da jede Suchmaschine SM_{Master} regelmäßig nicht mit allen Servern weltweit in Verbindung stehen wird. Nachdem der Informationssuchende sich für ein, ihn interessierendes Fachgebiet entschieden und die Suche nach einem ersten Deskriptor gestartet hat, läuft das Verfahren zur selektiven Informationsgewinnung in der vorstehend beschriebenen Weise.

Durch das selbstlernende System wird der zeitliche und finanzielle Aufwand für eine selektive Informationsrecherche signifikant verringert.

Bezugszeichenliste

AS Arbeitsspeicher
D Deskriptor
D_{unzul} unzulässiger Deskriptor
D_{syn} synonymer Deskriptor
DS Datensatz
$DS_{ähnl}$ Datensatz mit ähnlichem Deskriptor
DS_{red} redundanter Datensatz
DS_{rel} relevanter Datensatz
H(D) Häufigkeit des Deskriptors D
H(DS) Häufung des Datensatzes DS
I_D Informationsdichte
n_{DSrel} Anzahl relevanter Datensätze
n_{DSmax} maximale Anzahl relevanter Datensätze
$n_{ges.rel}$ Summe relevanter Datensätze
P_{rel} Wahrscheinlichkeit
S Server
SM Suchmaschine
SM_{Master} Master-Suchmaschine
SM_{Slave} Slave-Suchmaschine
SP Speicher
V_{DSrel} Volumen eines relevanten Datensatzes

Patentansprüche

1. Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken, insbesondere in regionalen und globalen Datennetzen, wobei ein relevantes Recherchesuchwort (Deskriptor D) über die Eingabeeinrichtung eines Computers eingegeben und über Informationsübertragungseinrichtungen die Verbindung mit einer Suchmaschine SM hergestellt wird,
wobei der Computer und/oder die Suchmaschine SM die Korrektheit und Zulässigkeit des Deskriptor D überprüft und bei unzulässigen Deskriptoren D_{unzul} synonyme Deskriptoren D_{syn} ermittelt und dem Nutzer vorschlägt,
daß nachfolgend alle, mit der Suchmaschine SM in Verbindung stehenden Server $S_1 \ldots S_n$ nach diesem Deskriptor abgefragt und von allen, auf den Servern $S_1 \ldots S_n$ ermittelten Datensätzen DS bruchstückartige Blöcke mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschine SM geladen werden,
daß nachfolgend redundante Datensätze DS_{red} eliminiert werden,
daß die Häufung $H(DS_i)$ der auf den einzelnen Servern $S_1 \ldots S_n$ ermittelten Datensätze DS_i verglichen und die Gesamtzahl $n_{ges.rel}$ der ermittelten, relevanten Datensätze DS_{rel} nach Eliminierung redundanter Datensätze DS_{red} bestimmt wird,
wobei die Recherche abgebrochen wird und die ermittelten Datensätze auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen werden, falls die Anzahl der ermittelten Datensätze n_{DSrel} kleiner als ein wählbares, vom Benutzer oder vom Rechercheprogramm vorgebbares Maximum n_{DSmax} ist oder
daß ein weiterer Recherchesuchlauf durchgeführt wird, falls die ermittelte Anzahl relevanter Datensätze n_{DSrel} größer als dieser Grenzwert ist,
wobei durch den Nutzer oder das Rechercheprogramm ein weiteres relevantes Suchwort vorgegeben und der zweite Recherchendurchlauf gestartet wird,
wobei die Server S_1 bis S_n von der Suchmaschine in der Reihenfolge der Häufung $H(DS_{rel})$ der Anzahl der ermittelten, relevanten Datensätze DS_{rel} abgefragt werden,
daß nachfolgend bruchstückartige Blöcke mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschine SM geladen werden,
daß nachfolgend redundante Datensätze DS_{red} gelöscht werden,
und die Recherche abgebrochen wird und die ermittelten Datensätze auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen werden, falls die Anzahl der ermittelten Datensätze n_{DSrel} kleiner als ein wählbares, vom Benutzer oder vom Rechercheprogramm vorgebbares Maximum n_{DSmax} ist oder
daß ein weiterer Recherchesuchlauf durchgeführt wird, bis die Forderung $n_{DSrel} <= n_{DSmax}$ erfüllt ist.

2. Verfahren zur Datenermittlung und -aufbereitung in Informationsnetzwerken, insbesondere in regionalen und globalen Datennetzen, wobei ein relevantes Recherchesuchwort (Deskriptor D) über die Eingabeeinrichtung eines Computers eingegeben und über Informationsübertragungseinrichtungen die Verbindung mit einer Suchmaschine SM_{Master} hergestellt wird,
die ihrerseits Verbindungen mit n weiteren Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ innerhalb des Netzes herstellt, wobei jede dieser n Suchmaschinen mit einer Anzahl von Servern ($S_{1.1}$, $S_{1.2}$, $S_{1.k} \ldots S_{i.n-2}$, $S_{i,n.1}$, $S_{i.n}$) in Verbindung steht,
daß nachfolgend alle, mit der Suchmaschine SM in Verbindung stehenden Server ($S_{1.1, S1.2}$, $S_{1.k} \ldots S_{i.n-2}$, $S_{i.n.1}$, $S_{.n}$) nach dem Deskriptor D abgefragt und von allen, auf den Servern ermittelten Datensätzen DS bruchstückartige Blöcke mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ geladen werden,
daß nachfolgend redundante Datensätze DS_{red} eliminiert werden,
daß auf jeder Suchmaschine $SM_{Slave.1}$ gespeichert wird, auf welchem der angewählten Server S_j welche Anzahl von Datensätzen ermittelt wurde,
daß nachfolgend ein Vergleich der Anzahl oder der Häufung der von den einzelnen Slave-Suchmaschinen $SM_{Slave.1} \ldots SM_{Slave.n}$ ermittelten Datensätze vorgenommen wird,
wobei redundante Datensätze ermittelt und ausgeschieden werden,
daß die Anzahl $n_{ges.rel}$ der ermittelten, relevanten Datensätze DS_{rel} ermittelt wird,
wobei die Recherche abgebrochen wird und die ermittelten Datensätze auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen werden, falls die Anzahl der ermittelten Datensätze n_{DSrel} kleiner als ein wählbares, vom Benutzer oder vom Rechercheprogramm vorgebbares Maximum n_{DSmax} ist oder
daß ein weiterer Recherchesuchlauf durchgeführt wird, falls die ermittelte Anzahl relevanter Datensätze n_{DSrel} größer als dieser Grenzwert ist,
wobei durch den Nutzer oder das Rechercheprogramm ein weiteres relevantes Suchwort vorgegeben und der zweite Recherchendurchlauf gestartet wird,
wobei die Server S_1 bis S_n von der Suchmaschine in der Reihenfolge der Häufung $H(DS_{rel})$ der Anzahl der ermittelten, relevanten Datensätze Ds_{rel} abgefragt werden,
daß nachfolgend bruchstückartige Blöcke mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschine SM geladen werden,
daß nachfolgend redundante Datensätze DS_{red} gelöscht werden,
und die Recherche abgebrochen wird und die ermittelten Datensätze auf den Arbeits- oder Massespeicher des Computers des Benutzers übertragen werden, falls die Anzahl der ermittelten Datensätze n_{DSrel} kleiner als ein wählbares, vom Benutzer oder vom Rechercheprogramm vorgebbares Maximum n_{DSmax} ist oder
daß ein weiterer Recherchesuchlauf durchgeführt wird, bis die Forderung $n_{DSrel} <= n_{DSmax}$ erfüllt ist.

3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet,
daß eine Korrelation zwischen dem Datenvolumen V_{DSrel} eines relevanten Datensatzes DS_{rel} und der Häufigkeit $H(D_i)$ des Auftretens des jeweils gesuchten Deskriptors Dj innerhalb dieses Datensatzes vorgenommen und daraus die Informationsdichte des relevanten Datensatzes DS_{rel} bestimmt wird,
und daß nur von den Datensätzen $DS_{rel.1} \ldots DS_{rel.n}$ bruchstückartige Blöcke mit Angabe der Fundstelle in den Arbeitsspeicher AS der Suchmaschine SM geladen werden, die eine vorgegebene minimale Informationsdichte I_D aufweisen.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß nach der Übertragung bruchstückartiger Blöcke aller, auf den Servern $S_1 \ldots S_n$ ermittelten Datensätzen DS in den Arbeitsspeicher AS der Suchmaschine SM ähnliche Datensätze $DS_{ähnl}$ eliminiert werden.

5. Verfahren nach einem der Ansprüche 1 bis 4, dadurch gekennzeichnet, daß die Ergebnisse der Recherchenabfragen auf einem Inhaltsspeicher der beteiligten Master- und/oder Slave-Suchmaschine abgelegt werden.

6. Verfahren nach einem der Ansprüche 1 bis 5, dadurch gekennzeichnet, daß die Suchmaschinen SM_{Master} über Kommunikationswege untereinander zu einem neuronalen Netz verbunden werden.

7. Verfahren nach einem der Ansprüche 2 bis 6, dadurch gekennzeichnet, daß die Recherchefrage parallel an mehrere Master-Suchmaschinen übertragen wird, die im Informationsnetzwerk autonom arbeiten oder miteinander verbunden sind.

- Leerseite -